Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 10:12:48
REFERENCE NO: 226

Author Names & Affiliations

- Ward Wheeler - American Museum of Natural History
- Melanie Hopkins - American Museum of Natural History
- Frank Burbrink - American Museum of Natural History
- Peter Whiteley - American Museum of Natural History
- Mordecai-Mark Mac Low - American Museum of Natural History
- George Amato - American Museum of Natural History
- Juan Montes - American Museum of Natural History

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Computational Science, Molecular Systematics, Comparative Genomics, Astrophysics, Paleontology, Anthropology

Title of Submission

NSF CI 2030 – American Museum of Natural History and Beyond

Abstract (maximum ~200 words).

The needs for and demands placed upon information technology at the American Museum of Natural History (AMNH) continue to present challenges to the AMNH research community. The Museum, a world-class scientific research institution, is actively engaged in research that draws on its collections, expeditions, and experiments conducted using the latest technology, as well as direct numerical simulations. AMNH scientists are involved in the most current topics and issues in science and require advanced computing facilities to support their research.

Currently, the AMNH Science Computer Cluster Facility is the major resource used to advance and support the Museum’s research and educational initiatives. These facilities long ago became overtaxed by the computational demands of our research, and are being pushed to their limits by the sizable computational and storage requirements for data analysis and archiving. A lack of high-performing computing (HPC) resources has led researchers to improvise solutions. This situation puts the Museum at a competitive disadvantage in terms of scientific productivity, responding to funding opportunities, and the ability to attract and retain talent. To address this issue, the Museum must design a flexible and adaptive campus-wide HPC system that meets the current and future needs of AMNH researchers and their collaborators.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

Anthropology
AMNH's Anthropology Division has recently begun to use tools more familiar in biological evolution and systematics. Curator Peter Whiteley's research employs an anthro-informatic approach, utilizing computational and phylogenetic tools to build on linguistic, cultural, and genetic variation databases and to test hypotheses about sociocultural and demographic evolution. Several linguistic and cultural databases have already been developed and analyzed; others are planned. For language data, initial concentration on two major world language families will be expanded eventually to include all such families, involving a large extension of data, storage requirements, and computational capacity. Databases to date include:
• Crow-Omaha kinship systems and social environments worldwide: 167 cases, 90 variables, 16 states
• Native North American sociocultural systems: 272 cases, 90 variables, 16 states
• Uto-Aztecan language databases: 40 cases, 92 variables, 4,080 entries
• Bantu language databases: 542 cases, 92 variables, ~49,000 sound sequences
• Bantu sociocultural database: 202 cases, 90 variables, 16 states
• Bantu genetic and genomic database: 109 populations, 2,149 individuals (mtDNA); 81 populations (ChrY)

Formal analysis has also been applied to dating and analyzing manuscript variations in an ethnographic journal of Western North America in 1776; ongoing transcription will include ~18 manuscripts, ca. 50,000 words each; comparative computational analysis of insertion, deletion, and replacement events will be difficult given our current resources.

Systematic Biology
AMNH's biological science Divisions are involved in a broad variety of research concerning the evolutionary relationships among a wide range of diverse creatures. The basic problem of reconstructing evolutionary trees is NP-hard. These researchers, led by Curator Ward Wheeler, are at the forefront of algorithmic innovation and combinatorial heuristics, and have developed a series of heuristics for the tree-alignment problem and network analysis. These ideas (and other existing approaches) have been implemented in AMNH's open source, freely available POY software, which was preceded by MALIGN, and is now being followed by the more general PhylogeneticComponentGraph (PCG).

Phylogenetic Analysis of DNA and other data using dynamic homology
POY is a phylogenetic analysis program that supports multiple data types (e.g., morphology, nucleotides, genes, etc.). POY can perform true sequence optimization and phylogeny inference simultaneously (i.e., input sequences need not be pre-aligned). Insertions, deletions, and rearrangements can then be included in the overall tree score (under Maximum Parsimony), or in the model (under Maximum Likelihood). A variety of heuristic algorithms have been developed for this purpose and are implemented in POY.

Phylogenetic analysis on Forests of Component Graph
There are two important scenarios that POY cannot accommodate: 1) "horizontal" evolution, via phylogenetic networks in bacterial and viral pathogens, and linguistic analysis via "borrowed" words; and 2) collections of general component graphs with insufficient information to create a single fully connected scenario. To deal with these limitations, we are currently developing PCG, which adds richness to the set of evolutionary solutions, but entails additional computational complexity.

The fundamental problems involved in evolutionary graphs are often NP-hard, requiring the implementation of parallelized heuristic solutions. AMNH scientists have been heavily involved in this aspect of computational research, but the increasing size and complexity of modern data-sets vastly exceed our capacity to analyze the data we generate.

Evolution and the Tree-of Life
Integrating genomic and phenotypic data is essential for estimating the tree of life that includes both extant and extinct species. These phylogenies then serve as the backbone which, when properly examined in an ecological and behavioral data context, are critical for understanding modes of speciation, diversification rates, adaptive radiation via ecological opportunity, community assembly, and how traits map to genes as full genomic data are produced. These complex and heavily assimilated research programs provide a comprehensive view on the origins of life and taxonomy, helping understand how genomic changes scale up to species diversification.

These types of studies require massively parallelized computer programs operating across thousands of high-speed cores with large amounts of memory and storage to accommodate the assembly of genomes, estimation of phylogenies from thousands of loci, and application of machine learning techniques to properly address the input from hundreds of variables to handle these comparative

phylogenomic-ecologically driven questions.

Paleontology
In paleontological research, computational demands come from two sources: the need to handle large datasets and the need to do multiple iterations of the same analysis. Analyses often include both sources.

Large datasets come in different forms. For example, AMNH researchers have worked with data from the, the publicly-accessible Paleobiology Database of organisms of all geological ages, including nearly 1.3 million fossil occurrences from over 175,000 collections. The database's size and organization allow paleobiologists to ask questions about diversification at large scales (temporal, spatial, and taxonomic) that were not previously possible—but at the cost of high computational demand. Another type of large dataset used regularly comprises multivariate morphological data, which includes information about many physical traits instead of just one or two. Computational time can grow exponentially with increasing numbers of traits and so far, fast algorithms remain limited. Because of this, analyses involving a high number of dimensions is well beyond the capacity of a desktop computer. Hopkins recently published a paper in which she had to identify a subset of dimensions to analyze because using the full dataset was too computationally demanding.

Astrophysics
AMNH's Astrophysics research group, led by Dr. Mordecai-Mark Mac Low, relies on computational modeling using gas dynamical and magnetohydrodynamical (MHD) simulations to study planet and star formation and interstellar gas structure, with applications to galactic winds and galaxy formation, with three strands of computation-intensive research:

1. The study of the evolution of the smallest dwarf galaxies, whose observed properties seem to disagree with standard cosmological model predictions. To understand whether these disagreements result from inadequate treatments of standard gas physics in numerical simulations, detailed models are used to resolve star formation down to the level of individual stars, studying the ways in which the most massive stars affect their surroundings, including through thermal energy and cosmic rays from supernovae, ionizing radiation, and stellar winds. The research tracks the elements produced by these stars over multiple generations of star formation.

2. Star cluster formation in a turbulent, supernova-driven, interstellar medium. To model this process, a software framework is being developed tying together an N-body gravitational dynamics code for stellar motions and evolution codes, and an adaptive mesh refinement MHD code including self-gravity, radiative transfer using ray tracing, and models for heating and cooling of the gas. This code is fully parallelized and currently optimized out to the thousand-processor scale.

3. Exploring how the structure of protoplanetary disks of gas and dust determines solid object formation in them. Mac Low and collaborators have proposed that glassy beads, called chondrules, which constitute over half the mass of the oldest meteorites, may have formed in thin, heated regions produced by magnetized turbulence in the gas disk. Further investigations will study these regions' structure and location, including the effects of realistic physical descriptions of the magnetic diffusivity, including not only classical Ohmic resistivity, but also ion-neutral drift and the Hall effect.

The Astrophysics research group makes extensive use of external computing resources, with approximately 3 million CPU-hours currently allocated on NSF-funded XSEDE resources, 1.8 million hours on a NASA High End Computing cluster, and 20 million hours on the Dutch national supercomputer.

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

Scientific research at AMNH is no longer confined to the Museum campus. AMNH research is national and global in nature. It is critical that its cyberinfrastructure also reaches beyond its walls. Therefore, by Spring 2018, AMNH will be connected to its regional research and education network provider, NYSERNet (New York State Education and Research Network), with a dedicated network connection linking AMNH to the higher education community in New York State and beyond, including to ORION (Ontario, Canada Research & Innovation Optical Network), CA*net (Canadian National Research & Education Network), ESnet (U.S. Department of Energy, Energy Sciences Network), MAX (Mid-Atlantic Crossroads), TWAREN (Taiwan Advanced Research & Education Network), NoX (Northern Crossroads—New

England), and MAGPI (Philadelphia GigaPoP). Initially, this connection will be made at 1 Gbps, but will grow to 10 Gbps and beyond as demand and new applications require. NYSERNet will also serve as a gateway to Internet 2, a connection AMNH will seek to implement in order to provide access to other research and higher education facilities nationally.

To make full use of these connections, the parts of the AMNH network serving its research and education users will need to be upgraded accordingly. In 2015, the network core was upgraded to support 10 Gbps connections to various network IDFs throughout the campus, with the capability to scale to 40–100 Gbps if needed. The next phase of the project will be to deploy new distribution and access layer switches capable of driving 10 Gbps or more in network transfer speeds to desktops and lab systems requiring this throughput. Additionally, AMNH seeks to expand high-speed wireless access into labs, classrooms, other research and education facilities, realizing that for many of our users, wireless is their primary connection method.

While these upgrades will improve the overall experience for the AMNH research and education community, even this infrastructure may not adequately address the needs of certain high-performance scientific applications. Therefore, AMNH will design and deploy a “Science DMZ”, a purpose-built high-speed network designed to facilitate high-volume bulk data transfer, remote experiment control, and data visualization. The Museum’s Science DMZ, to be called AMNH-SciNET, will be deployed alongside the existing campus network and will interconnect the scientific research departments, labs, imaging and visualization facilities (including the AMNH Microscopy and Imaging Facility [MIF]), the Hayden Planetarium), High Performance Computing (HPC) systems, and visualization clusters. To facilitate bulk file transfers between AMNH and collaborating institutions, a series of Data Transfer Nodes (DTNs) will be deployed within the AMNH-SciNet, possibly using the FIONA architecture developed at the University of California, San Diego.

AMNH’s IT Department has also begun the implementation of IPv6 throughout the Museum’s technology infrastructure. To that end, they have obtained an IPv6 allocation from ARIN (2620:0:2840::/48), which they are in the process of deploying to the core network infrastructure (border routers, core network switches, firewalls, DNS servers, etc.). In order to facilitate a seamless transition from IPv4 to IPv6, all network devices and capable endpoints will be configured in a dual-stack IPv4/IPv6 configuration. The second phase of the IPv6 rollout will focus on our research departments, and all advanced systems (HPC clusters, research specific networks, Science DMZs, WAN connections), and similar devices and services will be deployed as IPv6 capable beginning in 2018, with campus-wide IPv6 access being in place by 2020.

In order to ensure that network performance is maintained throughout the AMNH campus, and that the efficacy of upgrades and enhancements can be measured, AMNH will integrate the PerfSONAR infrastructure into both the AMNH-SciNET and the campus network to measure end-to end performance.

A final key component supporting high-speed computing throughout the infrastructure is data security. AMNH aims to provide data security in a sustainable manner that accounts for the varied risks and needs of a given system or data. AMNH has already made strides in segmenting its infrastructure into logical groupings based on common needs and security requirements. This segmentation will allow AMNH to apply the appropriate levels of security to those assets as needed. This means that the AMNH IT Department can apply stringent security controls on systems housing sensitive data (whether research data, administrative data, or otherwise) while maintaining appropriate levels of openness on less sensitive segments of the network.

The on-site computing clusters currently utilized by AMNH is unable to meet the many computationally-intensive research needs of the Museum’s large scientific staff. As the descriptions of the different research projects above indicate, AMNH is pioneering new approaches to problems in a wide range of disciplines, all of which rely increasingly on advanced computational capability. Research at AMNH, including the research described in this application, is also highly collaborative, both among different divisions within the Museum and externally with other institutions. For example, the Ordovician marine biodiversification research is a multi-institutional collaboration among invertebrate paleontologists at AMNH, the Natural History Museum at the University of Oslo, the Finnish Museum of Natural History, and the University of California, Berkeley. The anthropology research currently includes both the Divisions of Anthropology and Invertebrate Zoology at AMNH, along with secondary support from the AMNH Sackler Institute of Comparative Genomics, and with regular collaboration among postdocs, students, and interns from New York-area universities, including Columbia University. The astrophysicists working in Dr. Mac Low’s research group collaborate with AMNH colleagues in the Department of Earth and Planetary Sciences, and beyond the Museum with minority-serving colleges within the City University of New York; with Columbia University; and with international collaborators at institutions such as the Niels Bohr Institute in Copenhagen and the Institute for Theoretical Astrophysics at the University of Heidelberg.

In general, AMNH researchers rely on individual laboratory equipment, which typically ranges from multicore desktop machines to 32 and 64 core Intel and AMD “vanilla” boxes. There is one ten-year-old 128 CPU cluster (32 x 4 Intel CPU with Infiniband interconnect) for general use. This cluster is well-beyond the end of its expected service life and completely insufficient to meet the needs of the AMNH scientific

enterprise.

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

The needs for and demands placed upon information technology continue to present challenges to the research community. As a world class scientific and research institution, AMNH's needs are no different than those of other universities and research institutions that are similarly positioned. Behind the Museum's 45 permanent exhibition halls, more than 200 scientists work in over 100 laboratories in departments that house more than 33 million specimens and artifacts. Museum scientists are actively engaged in research that draw on its collections, expeditions, and on experiments conducted with the latest technology in state-of-the-art facilities. From creating computer models of colliding galaxies to sequencing DNA, from excavating dinosaur fossils in Mongolia to studying coral reef systems in the Bahamas, from discovering planets outside the solar system to collaborating with "descendant communities" to uncover the cultures of the past and the present, Museum scientists are involved in the most current topics and issues in science.

The Science Computer Cluster Facility is a major resource used to advance and support AMNH research and educational initiatives. The science clusters are used by museum research scientists, postdoctoral fellows, and graduate and undergraduate students, whose work relies heavily on high-end capability computing in areas of biology, genomics, astrophysics, and anthropology.

Data storage and archiving, data mining, and dissemination of the collected data are significant challenges faced by the Museum and similar entities. While AMNH maintains a central SAN of over 250 TB on its campus—shared across various departments and disciplines—a great deal more data is being stored by individual researchers within their own labs. The Museum expects the need for data storage to grow into petabytes of data soon. The overgrowing collection of scientific and observational data requires a system for cataloguing, preserving, and disseminating it for use both internally and with the wider scientific community, and the fact that much of this data is stored in a decentralized manner only compounds the problem.

To address this issue, AMNH is seeking to increase the amount of centralized storage available for its research community, both through the acquisition of additional storage capabilities as well as the migration of appropriate storage loads to the cloud. AMNH has also begun to study and catalogue the data currently in its collection, with the aim of unifying it using a Digital Asset Management (DAM) System.

AMNH scientists are at the forefront of developing and utilizing cutting-edge approaches in computing paradigms to address problems of broad application in the biological and physical sciences. For instance, researchers in Invertebrate Zoology have developed and implemented phylogenetic algorithms that are used by scientists around the world, while Astrophysics researchers, in collaborations with scientists world-wide, are using high-resolution numerical simulation techniques to bring life to the Hayden Planetarium Space Shows. Furthermore, as part of the Museum's core mission to educate, train, and disseminate information, the clusters are leveraged within AMNH's educational programs, in order to promote the significance of high-performance computing for science and engineering to society. AMNH is committed to developing the next generation of scientists and science educators through the continued support of innovative education programs such as BridgeUp STEM (http://www.amnh.org/learn-teach/adults/bridgeup-stem/) as well as the already mentioned RGGS and MAT programs. Because of the growing need for computational resources in modern research, the ability to provide these resources to AMNH students and adult learners will become a focal point for the institution.

For example, OpenSpace is new open source interactive data visualization software designed to visualize the entire known universe and portray our ongoing efforts to investigate the cosmos. Bringing the latest techniques from data visualization research to the general public, OpenSpace supports interactive presentation of dynamic data from observations, simulations, and space mission planning and operations. The software works on multiple operating systems with an extensible architecture powering high resolution tiled displays and planetarium domes, making use of the latest graphic card technologies for rapid data throughput. In addition, OpenSpace enables simultaneous connections across the globe creating opportunity for shared experiences among audiences worldwide.

It is critical that AMNH stay on top of emerging trends in research computing and the cyberinfrastructure supporting these endeavors. AMNH IT staff will continue to take part in regional research and education network meetings and working groups facilitated by NYSERNet and will expand its involvement at a national level with the Internet 2, InCommon, Open Science Grid, Educause, and other programs and consortia focused on developing a sustainable network infrastructure supporting science.

Consent Statement

- "I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to display it on a publically available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode)."